Practical Assessment, Research, and Evaluation

Volume 15 *Volume 15, 2010* Article 10

2010

Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that Should Matter

Tess Miller

Saad Chahine

Ruth A. Childs

Recommended Citation
Miller, Tess; Chahine, Saad; and Childs, Ruth A. (2010) "Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that Should Matter," *Practical Assessment, Research, and Evaluation*: Vol. 15 , Article 10.
DOI: https://doi.org/10.7275/dzm4-q558
Available at: https://scholarworks.umass.edu/pare/vol15/iss1/10

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 10, July 2010 ISSN 1531-7714

Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that *Should* Matter

Tess Miller, *University of Prince Edward Island*
Saad Chahine & Ruth A. Childs, *Ontario Institute for Studies in Education, University of Toronto*

This study illustrates the use of differential item functioning (DIF) and differential step functioning (DSF) analyses to detect differences in item difficulty that are related to experiences of examinees, such as their teachers' instructional practices, that are relevant to the knowledge, skill, or ability the test is intended to measure. This analysis is in contrast to the typical use of DIF or DSF to detect differences related to characteristics of examinees, such as gender, language, or cultural knowledge, that should be irrelevant. Using data from two forms of Ontario's Grade 9 Assessment of Mathematics, analyses were performed comparing groups of students defined by their teachers' instructional practices. All constructed-response items were tested for DIF using the Mantel Chi-Square, standardized Liu Agresti cumulative common log-odds ratio, and standardized Cox's noncentrality parameter. Items exhibiting moderate to large DIF were subsequently tested for DSF. In contrast to typical DIF or DSF analyses, which inform item development, these analyses have the potential to inform instructional practice.

Differential item functioning (DIF) analysis is typically used to identify test items that are differentially difficult for respondents who have the same level of knowledge, skill, or ability but differ in ways that should be irrelevant to their performance on the test (e.g., females vs. males; francophones vs. anglophones). Differential step functioning (DSF) analysis is an extension of DIF that examines whether groups differ at score levels within polytomously-scored items. Although DIF and DSF analyses are most often used to examine differences based on attributes that should be irrelevant to performance, these techniques can also be used to compare groups that differ in ways that should matter. For example, students whose mathematics teachers used inquiry-based instructional practices might be expected to perform better on items that require them to explain their problem-solving approach than students whose teachers did not.

The purpose of this study is to illustrate the use of DIF and DSF analyses to detect differences in item difficulty for groups of students taught by teachers with different instructional practices. In contrast to a typical DIF analysis comparing the performance of groups that differ in ways that should not be relevant to performance, the purpose of this analysis is *not* to identify potentially biased test items and suggest where items should be changed. Instead, results from analyses such as these have the potential to inform teachers' instructional practices. For this illustration, we use data from two forms of Ontario's Grade 9 Assessment of Mathematics to compare the performance of students whose teachers differ in their use of inquiry-based instruction.

Mathematics Education in Ontario

In 1999, Ontario's Ministry of Education introduced a new provincial mathematics curriculum for Grades 9 and 10. The curriculum specified that students should become proficient in "applying the steps of an inquiry/problem solving process" (Ontario Ministry of

Education, 1999, p. 39). The 2005 revision (Ontario Ministry of Education, 2005) provided even more detail, specifying that students should develop

> [the] use of planning skills – understanding the problem (e.g., formulating and interpreting the problem, making conjectures) [and] making a plan for solving the problem; use of processing skills – carrying out a plan (e.g., collecting data, questioning, testing, revising, modeling, solving, inferring, forming conclusions) [and] looking back at the solution (e.g., evaluating reasonableness, making convincing arguments, reasoning, justifying, proving, reflecting); [and] use of critical/creative thinking processes (e.g., problem solving, inquiry). (Ontario Ministry of Education, 2005, p. 20)

In this paper, we have chosen the term "inquiry-based instruction" to refer to pedagogical approaches that encourage students to define mathematical problems and plan solution strategies, in addition to solving mathematical problems. As Jarrett (1997) noted, the term has been used to refer to a range of practices in science and mathematics education, from "highly structured hands-on activities and 'cookbook' experiments" through "guided inquiry or the use of science kits" to "students … generating their own questions and investigations" (p. 3). She observes that the latter has the most elements of inquiry, but there may be times when the former is appropriate. Importantly Jarrett (1997), and others (see, for example, Clements, 1997), caution against assuming that every hands-on activity requires students to engage in inquiry.

Research by Airasian and Madaus (1983), Guthrie, Schafer, Von Secker, and Alban (2000), Grouws and Cebulla (2000), and Linn and Harnisch (1981) suggests that instruction affects students' opportunity to learn and test performance (for a review of the literature on the effects of instruction, including the use of manipulatives and technology, see Colker, Toyama, Trevisan, & Haertel, 2003). In the context of Grade 9 mathematics, students whose teachers frequently use inquiry-based instruction would be expected to have more opportunities to develop and explain their problem-solving approaches.

This study investigated whether inquiry-based instructional practices of teachers affect the difficulty of problem-solving items for their students. The Grade 9 Assessment of Mathematics, developed by Ontario's Education Quality and Accountability Office (EQAO), is based on Ontario's mathematics curriculum. As we noted earlier, the curriculum requires students to develop proficiency not only in solving mathematical problems, but also in defining the problems and developing solution strategies; these skills are more easily observed in the assessment's constructed-response items, which require students to explain their problem solving approaches, than on the multiple-choice items. The Grade 9 Assessment of Mathematics was accompanied by a questionnaire for mathematics teachers, which includes questions about their instructional practices. By linking students' responses to the constructed-response items on the assessment to their teachers' answers to the questions about their instructional practices, we were able to investigate whether students whose teachers reported using inquiry-based instructional practices found the items less difficult than students whose teachers did not. Because it was possible that these students might find *all* the constructed-response less difficult – something that DIF and DSF analyses would not be able to detect if the matching criterion were the total score on the constructed-response items – the total score on the multiple-choice items was used as the matching criterion in the DIF and DSF analyses. Inquiry-based instructional practices may also affect students' performance on the multiple-choice items, but we expect the effect to be less on those items because the multiple-choice items do not require students to explain their problem-solving approaches. These analyses, in effect, examine differences in performance on the constructed-response items that are beyond any differences in performance that might be found on the multiple-choice items.

Method

Data

The Grade 9 Assessment of Mathematics student performance data (n = 153,688) and corresponding Teacher Questionnaire (n = 4,919) for the 2005/2006 school year were obtained from EQAO. The Teacher Questionnaire contained 109 items exploring teachers' classroom practices. The researchers selected three items that related to inquiry-based instruction:

a) This past semester or year, how often did you have your Grade 9 mathematics students do each of the following … conduct mathematical investigations (e.g., to demonstrate the inquiry process)?

https://scholarworks.umass.edu/pare/vol15/iss1/10
DOI: https://doi.org/10.7275/dzm4-q558

b) How often did you have your Grade 9 mathematics students engage in activities related to the following achievement categories … Thinking [Thinking is defined in the curriculum as the use of planning skills, problem solving skills, and critical/creative thinking processes]?

c) This past semester or year, how often did you use the following tools and strategies in assessing your Grade 9 students' progress in mathematics … investigations of mathematical concepts?

These items were selected because they are related to inquiry-based instruction. Unfortunately, it is impossible to know where the reported investigations fall on Jarrett's (1997) continuum of inquiry-based instructional practices; however, these items were the best indicators available. Each of these items had five response options: never, seldom, sometimes, often, and very often.

The student file contained 36 items, 12 of which were constructed-response items; of these, six were scored on a scale of 1 to 4 and six were scored dichotomously (these were re-coded to scores of 1 and 2). The 12 items were associated with three tasks, each task having four items. Responses that were coded as illegible, irrelevant, off-topic or missing were assigned the lowest score (1). The 24 multiple-choice items were summed to create the matching variable for the DIF and DSF analyses.

EQAO develops eight forms of the Grade 9 Assessment of Mathematics each year for all combinations of the following characteristics: language (French or English), program (Applied or Academic), and administration date (Winter or Spring). This analysis used two forms from 2005/2006: English Academic Spring and English Applied Spring. These forms were chosen because of the larger numbers of English than French students and the larger number of students taking the test in the Spring than the Winter (the Winter administration is for students who took their mathematics course in the Fall semester; the Spring administration is for students taking the course in the Spring semester plus those taking a full-year course). The original data file contained 59,199 students who sat the English Academic Spring form and 25,944 students who sat the English Applied Spring form.

The Teacher Questionnaire data were matched to the student achievement data to create one file that identified the teachers' responses to the three questionnaire items and students' achievement scores on the 12 constructed-response items and summed multiple-choice scores. Although many teachers teach both Academic and Applied mathematics courses, the Teacher Questionnaire asked teachers to complete only one form; teachers indicated the program for which they were completing the form. When matching teachers' and students' responses, those students whose teachers did not fill out the form for their program were dropped, leaving 38,949 students for the English Academic Spring form and 17,353 students for the English Applied Spring form. In addition, 162 students from the Academic form and 341 from the Applied form who received a total score of zero on the multiple-choice items and 46 students from the Academic form and 204 students from the Applied who did not answer any of the constructed-response items were dropped from the file. Next, 528 students from the Academic form and 344 students from the Applied form were dropped because their teachers did not answer all three of the items about inquiry-based practices. The resulting files for English Academic Spring and English Applied Spring contained 38,259 students (1,556 teachers) and 16,464 students (947 teachers), respectively.

Finally, for the DIF analyses, we needed to create two groups of teachers based on their answers to the three questions related to inquiry and problem solving. Table 1 provides the means, standard deviations, and distributions for the teachers' responses to these three questions. These statistics were calculated across teachers, not across students; it is important to note that the number of students matched with a teacher varies.

As Table 1 shows, teachers' responses were negatively skewed; that is, more teachers selected "often" or "very often" than "never" or "seldom." For the DIF analyses, we created a reference group of teachers who reported that they "often" or "very often" engaged in all three inquiry-based instructional practices and a focal group of those who reported that they "never" or "seldom" engaged in these practices (teachers who "sometimes" engaged in these practices and those who responded inconsistently across the three questions were not included). Only these teachers and their students were retained for subsequent analyses. For the Academic form, the reference group consisted of 8,935 students (368 teachers) and the focal group contained 260 students (10 teachers). For the Applied form, the reference group had 3,746 students (208 teachers); the focal group had 234 students (14 teachers). The number of students per teacher ranged from 1 to 42.

Table 1: Means, Standard Deviations, and Frequencies of Teachers' Questionnaire Responses

			Score Distributions				
Question	*M*	*SD*	Never	Seldom	Sometimes	Often	Very Often
Academic (n = 1,556)							
Mathematical Investigations	3.38	0.88	1.9%	10.6%	45.3%	32.0%	10.3%
Problem Solving	4.08	0.77	0.3%	1.7%	19.2%	47.7%	31.1%
Assessments	3.27	0.85	2.6%	11.9%	48.0%	30.9%	6.6%
Applied (n = 947)							
Mathematical Investigations	3.34	0.89	2.2%	13.3%	41.5%	34.3%	8.7%
Problem Solving	3.82	0.83	0.1%	5.3%	28.7%	44.7%	21.2%
Assessments	3.31	0.88	3.0%	10.8%	46.1%	32.2%	7.9%

Software

There are many available software programs to conduct DIF analysis. The difference in software is often based on the mathematical algorithms that detect DIF. The two primary methods for calculating DIF are the Mantel Hansel (MH) non-parametric method and item response theory. The analyses in this study were performed using Penfield's (2007b) DIFAS 4.0 software program, which uses MH for DIF and DSF. For a detailed explanation of DIFAS and the mathematical algorithms, refer to Penfield (2007a, 2007b) and Penfield, Gattamorta, and Childs (2009).

Analyses

As described earlier, DIF analyses detect overall differences in difficulty for an item and DSF analyses detect differences at score levels within items (Penfield, 2007a; Penfield, Gattamorta, & Childs, 2009). DSF analyses can be performed for dichotomously-scored items, but do not provide any additional information beyond the DIF analyses. DSF analyses are most useful for polytomously-scored items, such as the constructed-response items included on the Grade 9 Assessment of Mathematics that were scored from 1 to 4. The DSF analyses implemented in DIFAS are based on a cumulative step function – that is, the comparisons are between students who achieve a particular score or higher (e.g., 3 and above) and those who do not (e.g., less than 3). These step differences may otherwise be hidden (e.g., effect estimators with opposite signs or magnitudes can obscure a large effect, the process of aggregating effect estimators may yield an overall large effect when resulting from the summation of smaller and possibly insignificant effects; Penfield, 2007a).

DIF analysis was used to detect the effects of teachers' instructional practices on students' performance on 12 items from each of the two forms of the Grade 9 Assessment of Mathematics. The Liu Agresti Cumulative Common Log-odds Ratio ($\hat{\alpha}_{LA}$), established by Penfield (2007a) as an equivalent metric to the Mantel-Haenszel and implemented in DIFAS 4.0 (Penfield, 2007b), was used to identify the effect size. Penfield (2007a) provided a classification scheme for categorizing the level of DIF in polytomous items where $|\hat{\alpha}_{LA}| < 0.53$ is negligible DIF, $0.53 \leq |\hat{\alpha}_{LA}| < 0.74$ is moderate DIF, and $|\hat{\alpha}_{LA}| \geq 0.74$ is considered large DIF.

Items having moderate or large effect sizes were further examined using three different DIF tests of significance, as recommended by Penfield (2007a): (1) the Mantel Chi-Square, which is distributed as chi-square with one degree of freedom (a critical value of 3.84 for an α of .05 was used in this analysis); (2) the standardized $\hat{\alpha}_{LA}$, which is the $\hat{\alpha}_{LA}$ divided by the estimated standard error (LOR Z; standardized $\hat{\alpha}_{LA}$ values greater than 1.96 or less than -1.96 may indicate the presence of DIF); and (3) the standardized Cox's noncentrality parameter (COX Z; values greater than 1.96 or less than -1.96 may indicate the presence of DIF).

Following the DIF analysis, a DSF analysis was performed on each item that had been found to exhibit DIF. This analysis had the potential to pinpoint, for example, at which score step the items were differentially difficult for the students whose teachers indicated they used different instructional practices. The DSF analysis generates three statistics: (1) weighted and

unweighted estimates of the DSF effect variance (CU-LOR; |CU-LOR| > 0.4 indicates a moderate effect and |CU-LOR| >0.6, a large effect); (2) standard error estimators of the weighted and unweighted estimates of the DSF effect variance (SE); and (3) the ratio of each DIF effect variance estimate divided by its standard error estimator (Z; Z statistics greater than 1.96 or less than -1.96 may indicate the presence of DSF) (Penfield, 2007a). The CU-LOR was first examined to determine the level of DSF in each step. Items with moderate or large effect sizes, as indicated by the CU-LOR, were further analyzed by examining Z. In these analyses, the number of multiple-choice items answered correctly was used to match students.

Results and Discussion

Students' Item Scores

Table 2 displays the means, standard deviations, and distributions for student responses in the reference and focal groups for the 12 items in each of the forms. It is important to note that each form had different items. To reflect the items' grouping into tasks and to differentiate the forms, we numbered the items C11-C14, C21-C24, and C31-C34 for the Academic form and P11-P14, P21-P24, and P31-P34 for the Applied form.

As Table 2 shows, the items varied widely in difficulty. More than half of the students taking each form earned the highest score of four on Items C22 and P23. For the Applied form, however, there were several items on which fewer than 10% of the students obtained a score of four. For both forms, the average item scores of the students whose teachers used inquiry-based instructional practices (the reference group) were higher than for the students whose teachers did not. The reference group also performed significantly better on the multiple-choice items for the Applied form: the average total score on the 24 multiple-choice items was 15.59 (SD = 4.14) for the reference group and 14.86 (SD = 4.60) for the focal group, $t(3978) = 2.59$, $p = .009$. For the Academic form, the difference in total multiple-choice score was not significant: the average was 15.30 (SD = 4.13) for the reference group and 14.80 (SD = 4.23) for the focal group, $t(9193) = 1.92$, $p = .055$. Because the total score on the multiple-choice items is used as the matching criterion in the DIF and DSF analyses, the analyses effectively are looking for differences in performance on the constructed-response items that are beyond the differences in performance on the multiple-choice items.

DIF and DSF Analyses

Table 3 provides the DIF analysis results. The last column indicates the direction of DIF, where inquiry represents teachers' responses of "often" or "very often" to the three questions about their practices.

Items with moderate or large DIF that were polytomously scored (i.e., Items C22, C24, and P34) were further analyzed for DSF. The findings are presented in Table 4. Note that Items C12 and C21 also showed moderate or large DIF but were not suitable for DSF analysis because they were dichotomously scored.

Item C12 was not available among the released items, but Items C22 and C24 were. The set of items containing these items (Items C21-C24) appears in Appendix A and is shown exactly as it appeared on the Grade 9 Assessment of Mathematics. All of the items associated with the "Choc-o-Can" task showed statistically significant DIF (although for Item C23, the size of the DIF was negligible).

Appendix B shows Item P34, which also had large DIF in favour of the reference group, along with its scoring guide. It is important to acknowledge that this was the last item presented in Booklet 2; hence it is possible that the DIF was due to a higher percentage of students in the focal group running out of time.

Although these analyses provide evidence of DIF in the set of items described as Choc-o-Can and in an item requiring students to calculate the area of an abstract shape (i.e., a sail), we must be cautious about concluding that teachers' inquiry-based instructional practices make a difference in students' performance on constructed-response items. Firstly, it would be necessary to compare the items that showed DIF and those that did not show DIF. Unfortunately, we were unable to compare the levels of cognition required to solve each item because many items in these assessments had not been released.

The second reservation stems from the DSF findings. The interpretation of DSF at the lower score points may be quite different from DSF at the higher score points. For example, the distinction among the lower scores may be conceptual understanding, while the difference between the higher scores may be minor errors of computation. We would expect students' conceptual understanding to be more affected by teachers' instructional practices. Unfortunately, our findings are not conclusive in this regard.

Table 2: Means, Standard Deviations, and Frequencies of Students' Item Scores

				Score Distributions			
Item	Group	*M*	*SD*	1	2	3	4
Academic ($n_{Reference}$ = 8,935, n_{Focal} = 260)							
C11	Reference	1.94	0.24	6.0%	94.0%		
	Focal	1.92	0.28	8.5%	91.5%		
C12	Reference	1.70	0.46	30.3%	69.7%		
	Focal	1.56	0.50	44.2%	55.8%		
C13	Reference	2.53	1.12	24.6%	23.7%	26.2%	25.5%
	Focal	2.38	1.12	30.0%	23.1%	26.2%	20.8%
C14	Reference	2.53	1.15	28.9%	13.8%	33.0%	24.3%
	Focal	2.40	1.19	33.8%	16.9%	24.2%	25.0%
C21	Reference	1.80	0.40	19.7%	80.3%		
	Focal	1.66	0.48	34.2%	65.8%		
C22	Reference	3.19	1.11	16.0%	6.4%	20.3%	57.2%
	Focal	2.72	1.33	33.5%	5.0%	17.3%	44.2%
C23	Reference	1.48	0.50	51.9%	48.1%		
	Focal	1.36	0.48	63.8%	36.2%		
C24	Reference	2.55	1.34	37.3%	10.6%	12.3%	39.8%
	Focal	2.08	1.32	54.6%	10.0%	7.7%	27.7%
C31	Reference	1.68	0.47	32.5%	67.5%		
	Focal	1.59	0.49	41.2%	58.5%		
C32	Reference	2.27	1.08	28.4%	35.9%	15.8%	19.9%
	Focal	2.05	1.07	38.8%	33.1%	12.3%	15.8%
C33	Reference	1.72	0.45	27.5%	72.5%		
	Focal	1.64	0.48	36.2%	63.8%		
C34	Reference	2.35	1.16	30.5%	30.0%	14.0%	25.5%
	Focal	2.04	1.16	45.8%	23.8%	11.2%	19.2%
Applied ($n_{Reference}$ = 3,746, n_{Focal} = 234)							
P11	Reference	1.53	0.50	47.0%	53.0%		
	Focal	1.52	0.50	48.3%	51.7%		
P12	Reference	1.36	0.48	64.3%	35.7%		
	Focal	1.31	0.46	68.8%	31.2%		
P13	Reference	1.65	0.88	54.1%	34.9%	2.8%	8.1%
	Focal	1.50	0.79	63.2%	29.5%	1.7%	5.6%
P14	Reference	1.70	1.09	65.3%	13.1%	7.8%	13.9%
	Focal	1.46	0.90	75.2%	11.1%	6.4%	7.3%
P21	Reference	1.82	0.38	17.8%	82.2%		
	Focal	1.82	0.39	17.9%	82.1%		
P22	Reference	1.89	0.31	10.8%	89.2%		
	Focal	1.84	0.37	15.8%	84.2%		
P23	Reference	3.16	1.21	19.3%	8.8%	9.0%	63.0%
	Focal	3.02	1.27	23.1%	10.3%	8.1%	58.5%
P24	Reference	2.48	1.29	36.9%	11.2%	18.5%	33.3%
	Focal	2.28	1.28	43.6%	12.8%	15.4%	28.2%
P31	Reference	1.34	0.47	66.1%	33.9%		
	Focal	1.28	0.45	72.2%	27.8%		
P32	Reference	1.31	0.46	69.1%	30.9%		
	Focal	1.29	0.45	71.4%	28.6%		
P33	Reference	1.99	1.00	45.0%	16.9%	32.4%	5.8%
	Focal	1.82	0.96	53.4%	14.5%	28.6%	3.4%
P34	Reference	1.55	0.93	68.0%	17.0%	7.4%	7.7%
	Focal	1.29	0.72	82.5%	10.7%	2.6%	4.3%

Table 3: Differential Item Functioning Results

Item	Effect Size α_{LA}	DIF Tests of Significance: Mantel	LOR Z	COX Z	DIF	Direction of DIF
Academic						
C11	0.325	1.954	1.395	1.397		
C12	0.608	20.212	4.406	4.500	Moderate	Inquiry
C13	0.212	3.359	1.843	1.840		
C14	0.149	1.616	1.252	1.262		
C21	0.769	29.670	5.267	5.444	Large	Inquiry
C22	0.765	41.494	6.120	6.439	Large	Inquiry
C23	0.519	12.981	3.555	3.601	Negligible	Inquiry
C24	0.793	33.034	5.664	5.754	Large	Inquiry
C31	0.348	6.832	2.578	2.614		
C32	0.355	8.919	2.910	2.986		
C33	0.387	6.953	2.651	2.638		
C34	0.501	16.689	3.976	4.088	Negligible	Inquiry
Applied						
P11	-0.104	0.413	-0.658	-0.641		
P12	0.125	0.578	0.772	0.760		
P13	0.327	4.109	2.019	2.024	Negligible	Inquiry
P14	0.515	8.852	3.029	2.981	Negligible	Inquiry
P21	-0.180	0.878	-0.942	-0.935		
P22	0.247	1.435	1.199	1.198		
P23	0.071	0.248	0.493	0.502		
P24	0.197	2.279	1.515	1.502		
P31	0.219	1.855	1.352	1.361		
P32	0.005	0.001	0.031	0.030		
P33	0.248	3.074	1.784	1.753		
P34	0.756	15.571	3.838	3.941	Large	Inquiry

Note. α_{LA} is Liu Agresti Cumulative Common Log-odds Ratio; Mantel is the Mantel Chi-Square; LOR Z is α_{LA} standardized; COR Z is Cox's noncentrality parameter standardized.

Table 4: Differential Step Functioning Results

Item	Step	CU-LOR	Z
C22	2	1.030	6.801
	3	0.800	5.552
	4	0.542	3.894
C24	2	0.847	5.560
	3	0.874	5.438
	4	0.647	3.909
P34	2	0.763	4.161
	3	0.853	3.143
	4	0.567	1.658

In addition, our ability to classify teachers based on their practices was limited. As with many questionnaires surveying teachers' practices or beliefs, the teacher questionnaire asked for self-reported practices related to Grade 9 Mathematics. Self-reporting can be problematic given that what teachers say they do and what practices teachers actually engage in can be different. Teachers may have labeled a wide range of activities as investigations. Furthermore, teachers may have varied in how they interpreted the response options – especially, the distinction between "seldom" and "sometimes." In future administrations of the teacher questionnaire, perhaps examples of frequency for these descriptors and definitions of terms such as investigations could be provided.

Finally, the choice of matching criterion may have affected the results. For these analyses, we assumed that the constructed-response items provided unique opportunities for students to demonstrate their skill in explaining their problem solving approaches, so that inquiry-based instructional practices might have a greater effect on students' performance on constructed-response items than on their performance on multiple-choice items. However, it is likely that

students' performance on both types of items benefits from these types of instructional practices (indeed, the students in Applied courses whose teachers reported using inquiry-based instructional practices had significantly higher multiple-choice scores than those students whose teachers did not). An in-depth analysis of the skills required by the questions could lead to a better selection of items for the criterion.

In sum, examining DIF and DSF due to teachers' inquiry-based instruction is complicated by the imprecision of teachers' self-reported practices, compounded by the small number of relevant teacher questionnaire items and the limited information available about the skills required for the students to answer the assessment items.

Conclusion

Recall that the purpose of this study was to illustrate a novel use of DIF and DSF analyses. As we conducted this study we often found ourselves reflecting about the capacity for the analysis to help us understand if differences that we would expect to have an effect on item difficulty do in fact have such an effect. The answer, we believe, is a cautious yes. Using DIF and DSF in an atypical way, as in this study, can reveal the impact of teaching practices on item difficulty, but more importantly, on students' opportunity to learn specific skills.

As this study illustrated, the definition of groups based on variables such as teachers' self-reported practices is more difficult than that based on variables such as gender. This study also illustrated the challenge of defining a subset of items that would not be expected to be affected by the relevant instructional differences. However, unlike typical DIF and DSF analyses which inform item development, these analyses have the potential to tell us which instructional practices make a difference.

References

Airasian, P., & Madaus, G. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement*, *20,* 103-118.

Clements, D. H. (1997). (Mis?)constructing constructivism. *Teaching Children Mathematics*, *4*, 198-200. Retrieved June 29, 2010, from http://investigations.terc.edu/library/bookpapers/mis_constructing.cfm .

Colker, A. M., Toyama, Y., Trevisan, M., & Haertel, G. (2003, April). *Literature review of instructional sensitivity and opportunity to learn.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Guthrie, T., Schafer, W., Von Secker, C., & Alban, T. (2000). Contributions of instructional practices to reading achievement in a statewide improvement program. *Journal of Educational Research, 93*, 211–225.

Grouws, D., & Cebulla, K. (2000). *Improving student achievement in mathematics*. Brussels: International Academy of Education. Retrieved July 17, 2008, from http://www.ibe.unesco.org/fileadmin/user_upload/archive/publications/EducationalPracticesSeriesPdf/prac04e.pdf.

Jarrett, D. (1997). *Inquiry strategies for science and mathematics learning: It's just good teaching.* Portland, OR: Northwest Regional Educational Laboratory. Retrieved June 29, 2010, from http://educationnorthwest.org.

Linn, R., & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 19*, 109–118.

Ontario Ministry of Education. (1999). *The Ontario curriculum, Grades 9 and 10: Mathematics*. Toronto, ON: Queen's Printer for Ontario.

Ontario Ministry of Education. (2005). *The Ontario curriculum, Grades 9 and 10: Mathematics, revised.* Toronto, ON: Queen's Printer for Ontario. Retrieved June 29, 2010, from http://www.edu.gov.on.ca/eng/curriculum/secondary/math910curr.pdf.

Penfield, R. D. (2007a). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20,* 335–355.

Penfield, R. D. (2007b). *DIFAS 4.0 user's manual.* Retrieved March 8, 2008, from http://www.education.miami.edu/facultysites/penfield/index.html.

Penfield, R., Gattamorta, K., & Childs, R. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38 – 49.

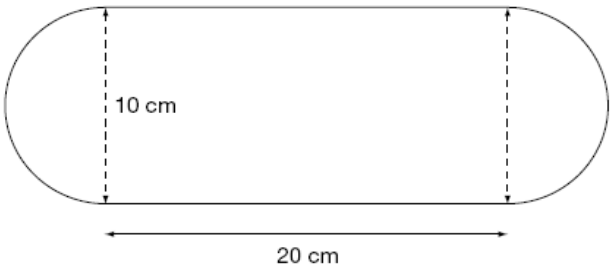
Appendix A

Items C21-C24 and scoring guide (in *Sample Assessment Questions and Scoring Guides*, available at www.eqao.com)

1. Choc-o-Can

Sweet Shapes is a company that makes chocolate. Each year, the company produces a new can for its specialty chocolates. This year's can is illustrated below. The top of the can swings open for easy access.

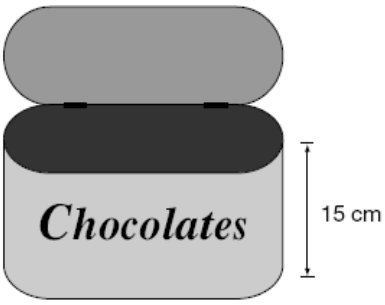
Derek makes a sketch of the bottom of the can and records the measurements below.



a) Determine the area of the bottom of the can.
Show your work.

b) The can contains individually wrapped chocolates that each take up about **28 cm^3** of space.

Determine how many chocolates a container of height 15 cm will hold.
Show your work.

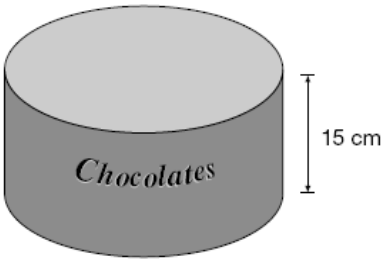


c) Sweet Shapes wants to reduce the size of each chocolate by 15%. Determine the volume of 100 of the reduced chocolates.
Show your work.

Reminder:
The original chocolates each take up about **28 cm^3** of space.

d) Next year, Sweet Shapes will produce a **cylindrical can** for the chocolates. The can will contain 75 wrapped chocolates, each with a volume of 19 cm^3. This can will also have a **height** of **15 cm.**

Determine the radius of this can.
Show your work.



Choc-o-Can (Spring)

B = Blank: nothing written or drawn in response to the question
I = • Illegible: cannot be read; completely crossed out/erased; not written in English
• Irrelevant content: does not attempt assigned question (e.g., comment on the task, drawings, "?", "!", "I don't know")
• Off topic: no relationship of written work to the question
U = Unacceptable
A = Acceptable

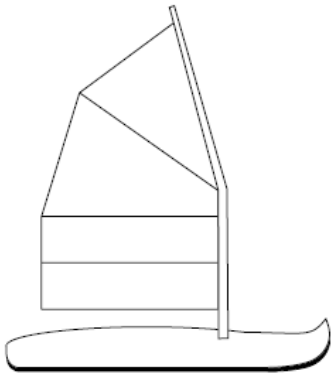
Part	Codes	Description
a)	U	
	A	Application of knowledge and skills to determine the area of the composite figure shows • an understanding of the concepts • an application of the procedures (e.g., $V = 278.5$ cm^2)
b)	10	Application of knowledge and skills to determine the number of chocolates shows limited effectiveness due to • misunderstanding of concepts • incorrect selection or misuse of procedures
	20	Application of knowledge and skills to determine the number of chocolates shows some effectiveness due to • partial understanding of the concepts • errors and/or omissions in the application of the procedures (e.g., correct number of chocolates, no work shown)
	30	Application of knowledge and skills to determine the number of chocolates shows considerable effectiveness due to • an understanding of most of the concepts • minor errors and/or omissions in the application of the procedures
	40	Application of knowledge and skills to determine the number of chocolates shows a high degree of effectiveness due to • a thorough understanding of the concepts • an accurate application of the procedures (any minor errors and/or omissions do not detract from the demonstration of a thorough understanding) (e.g., 149)
c)	U	
	A	Demonstration of understand of concepts of percent and procedures to calculate volume ($V = 2380$ cm^3)
d)	10	Problem-solving process to determine the radius shows limited effectiveness due to • minimal evidence of a solution process • limited identification of important elements of the problem • too much emphasis on unimportant elements of the problem • no conclusions presented • conclusion presented without supporting evidence
	20	Problem-solving process to determine the radius shows some effectiveness due to • an incomplete solution process • identification of some of the important elements of the problem • some understanding of the relationships between important elements of the problem • simple conclusions with little supporting evidence
	30	Problem-solving process to determine the radius shows considerable effectiveness due to • a solution process that is nearly complete • identification of most of the important elements of the problem • a considerable understanding of the relationships between important elements of the problem • appropriate conclusions with supporting evidence
	40	Problem-solving process to determine the radius shows a high degree of effectiveness due to • a complete solution process • identification of all important elements of the problem • a thorough understanding of the relationships between all of the important elements of the problem • appropriate conclusions with thorough and insightful supporting evidence (e.g., $r = 5.5$ cm)

Appendix B

Item P34 and scoring guide (in *Sample Assessment Questions and Scoring Guides*, available at www.eqao.com).

Marco is making the sail using green and red material in the ratio 3:2. He needs a total of 4.5 m^2 of material.

d) Determine how much **red** material he needs. Show your work.



d)	10	Application of knowledge and skills to determine the amount of red material, using ratios, shows limited effectiveness due to • misunderstanding of concepts • incorrect selection or misuse of procedures (e.g., does not multiply or divide, or uses numbers other than 2, 3, 5)
	20	Application of knowledge and skills to determine the amount of red material, using ratios, shows some effectiveness due to • partial understanding of the concepts • errors and/or omissions in the application of the procedures (e.g., multiplies or divides with one of 2, 3, 5)
	30	Application of knowledge and skills to determine the amount of red material, using ratios, shows considerable effectiveness due to • an understanding of most of the concepts • minor errors and/or omissions in the application of the procedures (e.g., multiplies and divides by wrong numbers [2, 3, 5])
	40	Application of knowledge and skills to determine the amount of red material, using ratios, shows a high degree of effectiveness due to • a thorough understanding of the concepts • an accurate application of the procedures (any minor errors and/or omissions do not detract from the demonstration of a thorough understanding) (e.g., multiplies by 2 and divides by 5 to get 1.8 m^2)

Note

The authors would like to thank the Education Quality and Accountability Office (EQAO) for providing the data for this research. The opinions expressed in this paper are solely those of the authors and do not necessarily reflect the opinions of EQAO. We would also like to thank Michael Kozlow and Randy Penfield for their advice on this study.

Citation

Miller, Tess, Chahine, Saad & Childs, Ruth A. (2010). Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that *Should* Matter. *Practical Assessment, Research & Evaluation*, 15(10). Available online: http://pareonline.net/getvn.asp?v=15&n=10.

Corresponding Author

Tess Miller
Memorial Hall 416
University of Prince Edward Island
550 University Avenue
Charlottetown, Canada C1A 4P3
TSMiller [at] upei.ca